

Representing genetic variation as continuous surfaces: an approach for identifying spatial dependency in landscape genetic studies

Melanie A. Murphy, Jeffrey S. Evans, Samuel A. Cushman and Andrew Storfer

M. A. Murphy (*mamurphy@wsu.edu*) and A. Storfer, *School of Biological Sciences, Dept of Biology, Washington State Univ., Pullman, WA 99164, USA.* – J. S. Evans, *USDA Forest Service, Rocky Mountain Research Station, 1221 S. Main, Moscow, ID 83483, USA.* – S. A. Cushman, *USDA Forest Service, Rocky Mountain Research Station, Missoula, MT 59801, USA.*

Landscape genetics, an emerging field integrating landscape ecology and population genetics, has great potential to influence our understanding of habitat connectivity and distribution of organisms. Whereas typical population genetics studies summarize gene flow as pairwise measures between sampling localities, landscape characteristics that influence population genetic connectivity are often continuously distributed in space. Thus, there are currently gaps in both the ability to analyze genotypic data in a continuous spatial context and our knowledge of expected of landscape genetic structure under varying conditions. We present a framework for generating continuous “genetic surfaces”, evaluate their statistical properties, and quantify statistical behavior of landscape genetic structure in a simple landscape. We simulated microsatellite genotypes under varying parameters (time since vicariance, migration, effective population size) and used ancestry (q) values from STRUCTURE to interpolate a genetic surface. Using a spatially adjusted Pearson’s correlation coefficient to test the significance of landscape variable(s) on genetic structure we were able to detect landscape genetic structure on a contemporary time scale (≥ 5 generations post vicariance, migration probability ≤ 0.10) even when population differentiation was minimal ($F_{ST} \geq 0.00015$). We show that genetic variation can be significantly correlated with geographic distance even when genetic structure is due to landscape variable(s), demonstrating the importance of testing landscape influence on genetic structure. Finally, we apply genetic surfacing to analyze an empirical dataset of black bears from northern Idaho USA. We find black bear genetic variation is a function of distance (autocorrelation) and habitat patch (spatial dependency), consistent with previous results indicating genetic variation was influenced by landscape by resistance. These results suggest genetic surfaces can be used to test competing hypotheses of the influence of landscape characteristics on genetic structure without delineation of categorical groups.

A key objective of ecological studies is to understand the influence of biotic and abiotic factors on population connectivity, and resulting fine-scale species’ distributions across a landscape. As a relatively new approach to assess connectivity, landscape genetics is an emerging discipline that aims to quantify the effect of landscape composition, configuration and matrix quality on the spatial distribution of genetic variation (Holderegger and Wagner 2006, Storfer et al. 2007). Evaluating neutral genetic variation in a landscape context has already provided insights into species’ ecology such as identification of potential barriers to gene flow (for review, see Manel et al. 2003, Storfer et al. 2007). However, our ability to quantify the relationship between genetic variation and multiple landscape variables with a robust assessment of error is currently constrained by the unique nature of multilocus data and available analytical tools (Storfer et al. 2007).

Compared to typical ecological data, spatial incorporation of neutral multilocus genetic data presents two unique complexities. Whereas typical ecological response variables (e.g. soil moisture, tree height, site occupancy) have direct

ecological interpretation and the observed value can be input directly into a statistical model, neutral genotypic measurements such as microsatellites are a collection of DNA fragment lengths. These data are only a meaningful response variable when being considered in the context of relationships among alleles, as allele frequency distributions, and/or differences in heterozygosity. In addition, ecological measurements can be associated with explicit spatial locations. In contrast, genetic summary statistics are generally not associated with sample locations, but represent an aspatial genetic distance between sample pairs (e.g. F_{ST} , Nei’s distance).

Several studies propose methods to estimate spatial location(s) of gene flow between pairs of sample locations (Arnaud 2003, Spear et al. 2005, McRae 2006, Cushman et al. 2006, McRae and Beier 2007). These approaches are extremely valuable; however, most methods only select a single path through the landscape based on ranked costs (but see McRae and Beier 2007). Although multiple costs or combinations can be tested (Cushman et al. 2006), empirical costs of movement through the landscape may be

unknown, difficult to hypothesize, or landscape variables may have a non-linear relationship with gene flow. In addition, significance of landscape genetic variation is often evaluated using the partial Mantel tests where significance tests may be unreliable (Raufaste and Rousset 2001, Rousset 2002, Castellano and Balletto 2002).

As an alternative, spatial autocorrelation statistics have been used to explain genetic variation in a spatially continuous manner (Slatkin and Arter 1991, Sokal et al. 1997, 1998, Epperson 2003, Shimatani and Takahashi 2003). However as opposed to strict spatial autocorrelation, the pattern of genetic variation may be the result of spatial dependency (Wagner and Fortin 2005), where another spatial variable drives genetic variation (e.g. habitat patch; Bockelmann et al. 2003, Keyghobadi et al. 2005). If this variable is autocorrelated, the result can be a significant autocorrelation statistic, deemed “false autocorrelation” (Legendre et al. 2002). In addition, there may be an interaction between autocorrelation and spatial dependency (Legendre et al. 2002, Fortin and Dale 2005). An optimal solution would allow testing of multiple landscape characteristics with competing hypotheses of autocorrelation, spatial dependency, and autocorrelation-spatial dependency interaction without the necessity of defining costs of independent variables (Storfer et al. 2007). To achieve this goal, raw multilocus genotypic data could be converted into point representations of genetic variation, creating a common theoretical base for data analysis between population genetics and spatial statistics (Shimatani and Takahashi 2003).

Representing whole genotypes as spatially referenced, continuous measures of genetic structure would advance our ability to test the effect of continuous landscape variables with well-developed multivariate spatial statistics, such as spatial regression methods (Fotheringham et al. 2002, Haining 2003) and point pattern analysis (Diggle 2003). These methods offer several advantages relative to widely applied methods in landscape genetics including: ability to estimate the influence of multiple independent variables simultaneously, parameter estimates for these variables that are valid in the presence of spatial autocorrelation, robust assessment of uncertainty, and ability to use landscape genetic models for spatial prediction (Wagner and Fortin 2005). Thus far, such methods have yet to be derived for landscape genetics.

To address these issues, we develop and evaluate an approach based on a novel integration of available methodological components for creating a continuous surface of genetic variation. We use a Bayesian clustering algorithm (Pritchard et al. 2000) to generate an ancestry value for each spatially referenced genotype to generate point values for surface interpolation. Via simulation, we address the following questions using a genetic surfacing approach: 1) does landscape genetics have the power to detect the effect of contemporary landscape condition? 2) How much data are needed to have the power to detect landscape genetic structure? 3) Can landscape genetics accurately identify the process(es) generating landscape genetic structure? 4) What is gained using a landscape genetic approach compared to standard global statistics (F_{ST})? Finally, we use a black bear genetic dataset from northern Idaho, USA (Cushman et al.

2006) to give an example of genetic surfacing with an empirical dataset.

Methods

Simulation components

Our genetic surfacing simulations had six main steps: defining the landscape, simulation of multilocus genotypes, subsampling genotypes, estimating ancestry, surface interpolation, and testing models of landscape genetic structure. We outline the steps of the simulations below and as a flow chart (Fig. 1). Additional details are located in Supplementary material, Appendix S1.

Step 1 – defining the Landscape

We defined two landscapes (48×48 km) in ArcInfo (ESRI 2005). Landscape A (Fig. 1A) contained two patches of equally suitable habitat separated by unsuitable habitat and was used to simulate a landscape effect independent of distance (i.e. spatial dependency). In landscape B (Fig. 1B), each patch was partitioned into five subpatches of equally suitable habitat for simulations including autocorrelation. In an empirical landscape, habitat patches could be defined by landscape variables (e.g. slope or moisture) or ideally, the landscape could be represented by continuous values (McGarigal and Cushman 2005).

Step 2 – simulation of multilocus genotypes

To address if landscape genetics has the power to detect the effect of current landscape condition, we simulated microsatellite genotypes in EasyPop 1.8 (Balloux 2001) under various conditions (Table 1). In EasyPop we parameterized either two (Fig. 1A; one “population” per habitat patch) or 10 “populations” (Fig. 1B; one “population” per subpatch). It should be noted that these are “populations” only as the population parameter in EasyPop (Balloux 2001). That is, with high levels of migration, they may be genetically indistinct. In landscape A, we varied time since vicariance (T , number of generations of separation), migration (M , probability per individual of migration per generation), and effective population size (N , number of breeding individuals contributing to the next generation) (Balloux 2001). In landscape B, we implemented a hierarchical stepping-stone model of migration (HM) (Balloux 2001). The hierarchical stepping-stone model of migration has two migration parameters: migration rate between subpatches within the same patch and migration rate between patches (Fig. 1B). We used this model to simulate a distance effect (i.e. autocorrelation without an effect of the unsuitable habitat) and an interaction between distance and landscape patch (Fig. 1B, see Table 1 for range of conditions). To model genetic variation within populations of a mobile species, we incorporated the potential for individuals to move freely within their given habitat patch within a generation. We implemented this by randomly assigning individuals within their patch (landscape A, Fig. 1) or subpatch (landscape B, Fig. 1) using ArcMap 9.1 (ESRI 2005) for each simulation.

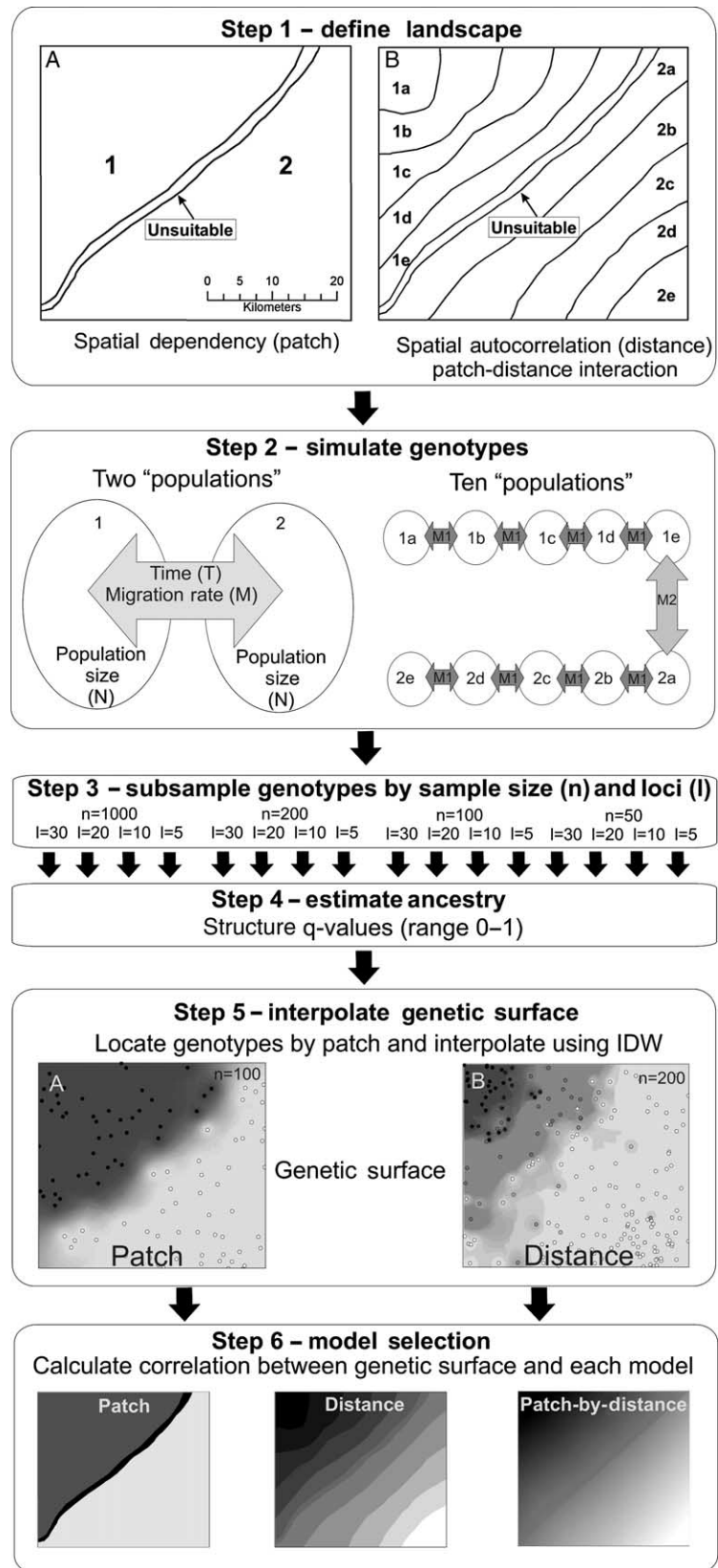


Figure 1. (Continued).

Table 1. Simulation parameters and power to detect landscape genetic structure with full dataset. The table displays EasyPop (Balloux 2001) simulation conditions as follows: simulation identification (Simulation), the number of generations post vicariance (Gen), probability of migration per individual per generation across the unsuitable habitat (Fig. 1) (M Between), probability of migration between adjacent subpatches within populations (Fig. 1) (M Within), total effective population size for the entire landscape (Total N_e), number of simulated (not genetically identified) populations or subpopulations (Pops), simulated model of genetic structure (MGS), and landscape used for simulation (Land). Simulated MGS was distance (Dist), habitat patch, or patch-distance interaction (PXD). We held all other parameters constant across simulations (see text). The table displays mean correlation coefficient (μR), number significant correlations (Sig, X/5), and standard deviation (SD) for correlations between the model of landscape genetic structure and genetic surface. Simulation codes are as follows: T – time since vicariance, M – migration rate, and HM – hierarchical migration model. For population size (N) and unequal effective population size (UN), see Supplementary material, Appendix S1b.

Simulation	Gen	M Between	M Within	Total N_e	Pops	MGS	Land	μR	Sig	SD
T1	0	0	Na	1000	2	None	A	0.071	0	0.027
T2	1	0	Na	1000	2	Patch	A	0.073	0	0.030
T3	5	0	Na	1000	2	Patch	A	0.542	5	0.225
T4	10	0	Na	1000	2	Patch	A	0.867	5	0.048
T5	50	0	Na	1000	2	Patch	A	0.986	5	0.003
T6	100	0	Na	1000	2	Patch	A	0.991	5	0.001
T7	250	0	Na	1000	2	Patch	A	0.991	5	0.001
T8	500	0	Na	1000	2	Patch	A	0.991	5	0.001
T9	1000	0	Na	1000	2	Patch	A	0.992	5	0.002
M1	500	0	Na	1000	2	Patch	A	0.991	5	0.001
M2	500	0.001	Na	1000	2	Patch	A	0.912	5	0.167
M3	500	0.005	Na	1000	2	Patch	A	0.990	5	0.002
M4	500	0.01	Na	1000	2	Patch	A	0.976	5	0.002
M5	500	0.05	Na	1000	2	Patch	A	0.949	5	0.014
M6	500	0.1	Na	1000	2	Patch	A	0.574	5	0.107
M7	500	0.2	Na	1000	2	Patch	A	0.175	1	0.120
HM1	500	0.001	0.05	1000	10	PXD	B	0.991	5	0.002
HM2	500	0.001	0.1	1000	10	PXD	B	0.991	5	0.002
HM3	500	0.001	0.2	1000	10	Patch	B	0.989	5	0.001
HM4	500	0.01	0.01	1000	10	Dist	B	0.866	4	0.123
HM5	500	0.01	0.05	1000	10	PXD	B	0.899	5	0.016
HM6	500	0.01	0.1	1000	10	PXD	B	0.976	5	0.012
HM7	500	0.01	0.2	1000	10	Patch	B	0.982	5	0.004
HM8	500	0.1	0.1	1000	10	Dist	B	0.897	5	0.020

Step 3 – subsampling

To address how much data are needed to detect landscape genetic structure, we varied sample size and number of loci used by subsampling genotypes and loci of each simulated dataset. Using R statistical package (R Development Core Team 2006), we randomly sampled genotypes without replacement (at 100, 20, 10, and 5%) across the entire landscape to emulate different levels of field sampling without knowledge of habitat patch or population boundaries. We then analyzed each subsample for four quantities of loci: all (30), 20, 10 and 5 loci.

Step 4 – estimating proportion of ancestry values

We used the Bayesian clustering algorithm STRUCTURE (Pritchard et al. 2000) to derive a continuous measure of genetic variation based on unclassified ancestry values (ranging from 0 to 1). STRUCTURE ancestry values are a proportion of ancestry of each individual's genotype to each of K populations ("genetic clusters"). A gradient in

genetic structure is represented by expressing the degree of ancestry in a given genetic cluster, maintaining the spatial and statistical variability within cluster. This is akin to fuzzy set theory as applied in remote sensing applications (Bosserman and Ragade 1982, Metternicht 2003), where the response variable may be proportion or probability of membership in multiple habitat classes. The number of clusters identified by STRUCTURE was selected using ΔK (Pritchard et al. 2000, Evanno et al. 2005), and appending ancestry values to each genotype in R (Fig. 1).

Step 5 – interpolation of a genetic surface

We interpolated a genetic surface from the ancestry values for each observed genotype using analyses that are comparable, repeatable, and automated across conditions and replicates. In no cases did STRUCTURE identify more than two clusters (Evanno et al. 2005); therefore, one genetic surface could be generated from the ancestry values for each individual for one of the two genetic clusters. In

Figure 1. Flow chart of simulation and genetic surfacing methodology. Step 1 – define landscapes. Both landscapes consist of suitable and unsuitable habitat. Simulations in landscape A contain spatial dependency only, while simulations in landscape B contain autocorrelation or an interaction between spatial dependency and autocorrelation. Step 2 – simulate genetic data. Circles represent "populations" in EasyPop (labels correspond to habitat patches or subpatches on the corresponding landscape). Simulations in landscape B had two migrations rates labeled as M1 (between habitat patches) and M2 (migration rate within habitat patches). Step 3 – subsample simulated data prior to analysis by sample size (n) and number of loci (l). Step 4 – analysis in program STRUCTURE. Step 5 – interpolate genetic surface. Patch and distance on interpolated surfaces indicate condition under which data were simulated and n is the sample sized used to create that particular genetic surface. Step 6 – model selection. We correlated each genetic surface with each model of landscape genetic structure (patch, distance, patch-distance interaction), tested significance, and selected model based on highest correlation coefficient. See Methods for additional details for each step.

addition, to quantify whether random effects can be correctly identified with a genetic surface (i.e. no structure), surfaces were interpolated from ancestry values for all data sets ($K=2$). We created genetic surfaces with an inverse distance weighted interpolation (IDW) (Cressie 1993) with the following parameters: a power function of 2 to control the tension of the surface, a variable search radius of 12 neighboring observations, and a resolution of 30 m².

Step 6 – test landscape genetic models

Under our simulated conditions, genetic variation could be due to spatial dependency (“habitat patch”), spatial autocorrelation (“distance”), both (“patch distance interaction”) (Legendre et al. 2002), or random effects. We calculated a raster to represent each hypotheses of landscape genetic structure in ArcInfo (ESRI 2002). The habitat patch raster consisted of a unique value assigned to each habitat patch (Fig. 1). The distance raster was mean distance (m) for a subpatch from patch 1a as a reference point (Fig. 1). The patch-distance interaction raster evaluated an interaction between habitat and distance. We achieved this by multiplying distance from the unsuitable habitat by patch value (1 for patch 1 and -1 for patch 2, Fig. 1). Using a modified version of program MODTTEST (Legendre 2000), we calculated a Pearson’s correlation coefficient between the genetic surface and each of the three models of landscape genetic structure under each simulation condition ($n=1000$) (Dutilleul 1993, Fortin and Payette 2002, Legendre et al. 2002). As implemented in MODTTEST, in our case the Pearson’s correlation is insensitive to violations of bivariate normal in the landscape patch analysis ($r=0.987$ to Spearman-rank correlation).

Power, accuracy and process identification

We assessed model performance by power, accuracy, and identification of the simulated process (distance, habitat patch, patch-distance interaction). We assessed statistical power ($1-\beta$) for both the complete simulated dataset and subsampling levels from step 3. This addresses both the power to detect landscape genetic structure for a given set of conditions, as well as the amount of data (sample size and number of loci) required to detect that landscape genetic structure.

We evaluated accuracy by two methods: presence of type I errors and correct identification of simulated landscape genetic structure (distance, habitat patch, or patch-distance interaction). Type I errors occur where significant genetic structure is detected in the absence of simulated genetic structure (Zar 1999, Legendre et al. 2002). We assessed type I error under two conditions: 1) genotypes with randomized xy coordinates and 2) simulation conditions where no landscape genetic structure was present. To assess the first, we randomized genotypes in step 2 and then completed the analysis twenty times to evaluate a condition known to have no significant landscape genetic structure (Fig. 1). For the second assessment of type I error, we identified simulations where the null hypothesis (no landscape genetic structure) was true (T1; simulations where no significant landscape genetic structure was detected using the complete dataset, Table 1). If landscape genetic

structure is not detected with the complete dataset, any subsample with a significant result is a type I error.

To evaluate the accuracy of variable selection, we calculated the number of instances where the simulated (“correct”) model of landscape genetic structure was identified out of the total number of analyses. As a general estimate of model choice, we calculated the number of instances where the “correct” relationship had the highest correlation coefficient out of the total number of significant analyses. For simulations in landscape A, the “correct” relationship was the habitat patch model (Table 1). For simulations in landscape B, the “correct” relationship was the distance model for HM4 and HM8 where migration was equal across the landscape (Table 1).

For the remaining simulations, identifying the “correct” relationship was more complex because we simulated multiple relationships simultaneously. We identified the “best” relationship based on the ability to detect landscape genetic structure in simulations from landscape A (Supplementary material, Appendix S1). For example, if the within patch migration probability is 0.2 and between patch migration 0.1, we would not expect to detect a distance effect if 0.2 probability of migration was undetectable in landscape A. Therefore, the “best” relationship in these simulations would be the habitat patch model. In addition, when the migration probability was disparate between versus within habitat patches, genetic surfacing may only detect the stronger restriction to migration probability. We considered the strongest correlation non-spurious as long as it identified either the “best” relationship or the landscape genetic model corresponding to the lowest migration probability. Then we evaluated the power to detect a significant interaction between autocorrelation and spatial dependency (i.e. patch-distance interaction) for different levels of migration within and between patches.

Comparison to standard genetic statistics

To evaluate the comparability of the genetic surfacing approach to standard genetic statistics, we calculated F_{ST} (standard measure of genetic distance based on heterozygosity) and allelic richness by patch for all simulations. We then calculated the number of significant genetic surfaces by level of genetic differentiation (F_{ST}) across all simulations.

Empirical illustration

We utilized a previously published empirical dataset of black bear *Ursus americanus* genotypes from northern Idaho, USA (Cushman et al. 2006) for a simple illustration of the genetic surfacing approach. The dataset consists of 146 unique black bear genotypes generated using nine microsatellite loci collected on a 2.6 km² grid (Cushman et al. 2006). We applied the genetic surfacing framework, with minor adjustments. So that results would be comparable with our simulations, we identified relatively coarse-scale habitat patches based on the major elevation break in a digital elevation model (DEM) for the study area (elevation = 800 m) determined by presence of suitable habitat and bear observations. This created four habitat

patches (Fig. 2): suitable habitat (1, 2, 4) and unsuitable habitat (3). Patch four (Fig. 2) was outside the study area; therefore it was excluded from the analysis. We used the resulting patch configuration to create the habitat patch, distance and patch-distance interaction models as in the simulation (Fig. 1, step 6). We executed STRUCTURE, interpolated the genetic surface, and performed the spatially adjusted Pearson's correlation as in the analyses of the simulated data (Fig. 1). We then compared the results to those of Cushman et al. (2006).

Results

Using genetic surfacing, we were able to detect landscape genetic structure generated by contemporary landscape condition (≥ 5 generations post vicariance, Table 1), with low to moderately high levels of migration (migration probability 0.0–0.1) and when population differentiation in minimal ($F_{ST} \geq 0.00015$). In addition, we demonstrate that the correlation between genetic variation and distance can be significant when the simulated process was habitat patch only (i.e. spatial dependency). However, by comparing multiple models (habitat patch, distance, patch-distance interaction) we were able to correctly identify the simulated relationship (e.g. habitat patch alone) 96% of the time (Supplementary material, Appendix S2).

We found genetic surfacing to be insensitive to misidentification of number of genetic clusters (K) identified in STRUCTURE. In all simulations, analyses in STRUCTURE suggested that either one or two genetic clusters (K) were most likely. However, in 9.37% (322/3434) of analyses, K = 2 was not a significant improvement over K = 1 (Table 1, Supplementary material, Appendix S2). Even still, we found evidence of continuous genetic structure. We identified significant landscape genetic structure for 16.15% (52/322) of these analyses with moderately high power (0.7453), and low type I error (3.73%, 12/322). In addition, when ancestry values for K

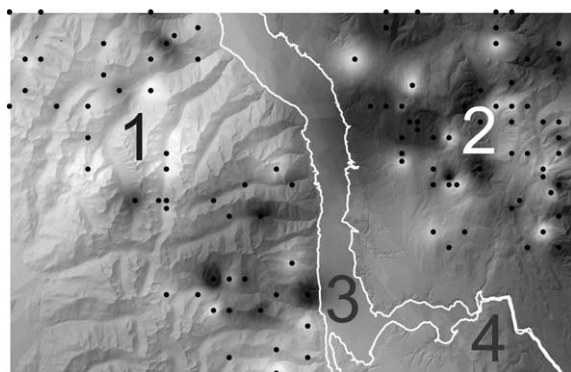


Figure 2. Empirical landscape – north Idaho, USA. The landscape was classified into habitat patches based on suitability as black bear habitat. Patch boundaries are superimposed on a shaded relief. Patches are labeled on the figure – 1, 3 are suitable habitat while 2 is unsuitable habitat. Patch 4 was outside the study area and therefore excluded from the analysis. Points represent sample locations overlaid on the genetic surface and shaded relief. Light areas represent high ancestry values while dark areas represent low ancestry values.

greater than two were used for surface interpolation for a cross section of simulations, Pearson's correlations and model choice did not significantly change ($r = 0.91$).

Statistical power – time, migration, and effective population size

Using genetic surfacing, we were able to detect the effect of current landscape condition five generations post vicariance, 0–0.1 probability of migration, for the range of simulated population sizes, and in the presence of internal patch structure (Table 1, Supplementary material, Appendix S1). The strength of the relationship between the landscape model and the genetic surface increased as time since vicariance increased from zero, significant for all simulations 5–1000 generations post vicariance (Table 1, Fig. 3A, B). Genetic surfaces appeared to be less distinct in the presence of migration between habitat patches (Fig. 3C, D). However, all analyses detected significant landscape genetic structure for simulations with migration probabilities between 0 (M1) and 0.1 (M6) ($r = 0.574$; Table 1). Variation in effective population size or unequal population size had little effect on the power to detect the simulated genetic structure, unless global effective population size was small (50) compounded by the presence of migration (Supplementary material, Appendix S1, S2). In the presence of internal patch structure (Fig. 1B), we were able to detect the simulated landscape genetic structure 97.5% of the time (Table 1).

Statistical power – data requirements

We gained more power by increasing sample size as opposed to increasing number of loci (Fig. 4a), with data requirements dependent on the amount of genetic structure. As we increased time since vicariance, fewer samples were necessary to have sufficient power to detect significant landscape genetic structure (Fig. 4b). For example, with 20% sampling genetic surfacing had 0.90 power 50 generations post vicariance but only 0.20 power to detect landscape genetic structure 10 generations post vicariance (Fig. 4c). At the same sampling level, when migration probability increased from 0.05 to 0.1, power decreased from 0.92 to 0.10 (Supplementary material, Appendix S2). However, power increased to 0.55 with complete sampling of simulations with migration probability of 0.1 (Fig. 4d). For more moderate levels of migration (< 0.005), we were able to detect landscape genetic structure regardless of sample size or number of loci (Fig. 4d, Supplementary material, Appendix S2). In general, population size and unequal population size had little effect on power (for details, see Supplementary material, Appendix S1, S2).

Accuracy and process identification

We were able to identify the presence of landscape genetic structure and the model used to simulate landscape genetic structure accurately. In the randomization assessment of accuracy, only 3.8% (468/12 300) of the correlations between the genetic surface and model of landscape genetic

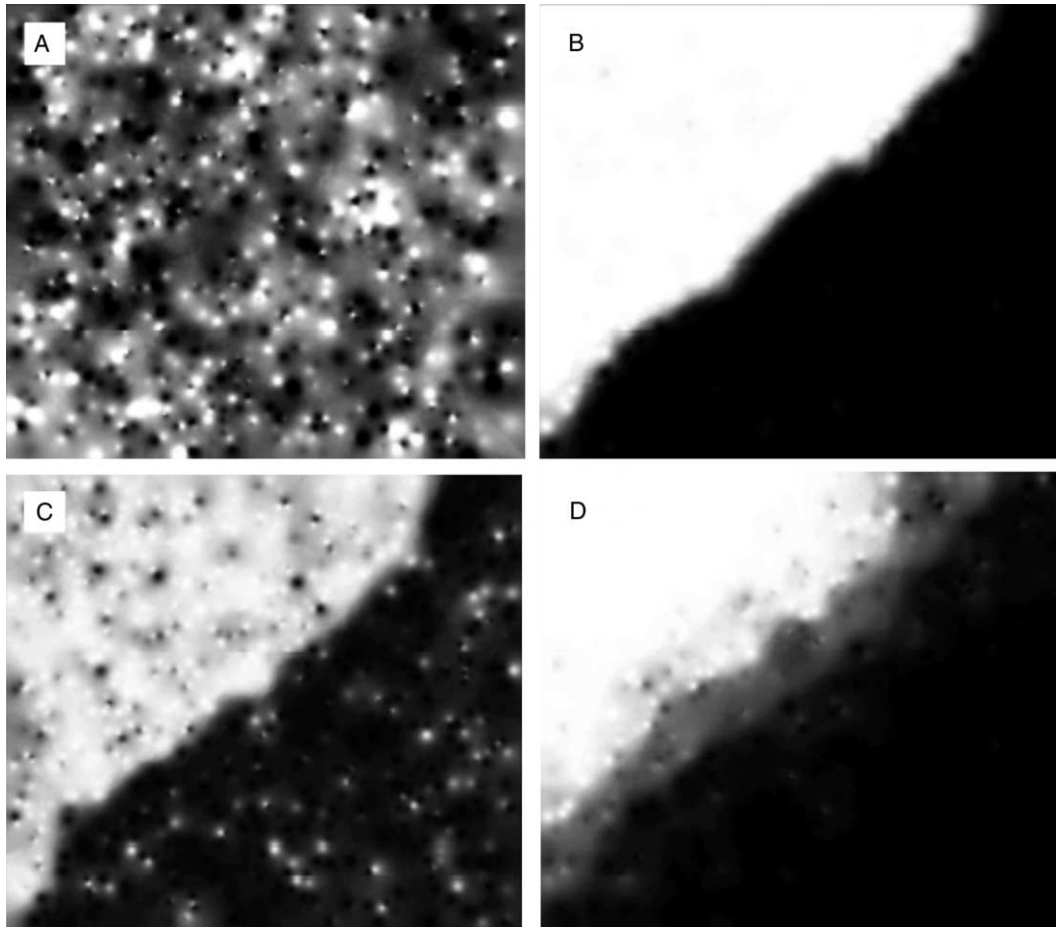


Figure 3. Example genetic surfaces interpolated from STRUCTURE ancestry values including all samples and loci (see methods for details). Light areas represent high ancestry values while dark areas represent low ancestry values. (A) No landscape genetic structure (T1: complete admixture); (B) strong differentiation in relation to habitat patch (spatial dependency; T8: 500 generations post vicariance); (C) with migration, genetic surfaces become more variable but are still differentiated (M4; 0.01 between patch migration probability); (D) internal patch structure results in a gradient of ancestry values across the landscape (HM4; 0.01 between patch and within patch migration probability).

structure were significant, below the selected p-value (0.05). The type I error rate was also below 5% for all samples sizes and number of loci (0.11–0.48). In simulation conditions without landscape genetic structure (T1, T2, M7), the type I error was 5% (12/240). When evaluating the accuracy to identify the simulated process (habitat patch, distance, patch-distance interaction), all correlations between the genetic surface and each model of landscape genetic structure was often significant (Fig. 5 a, b, c). After model comparison the “correct” model was selected for most simulations without internal patch structure (>95%; Supplementary material, Appendix S2, Fig. 5d, e). When we simulated a distinct hierarchical effect (HM5), genetic surfacing selected the patch-distance interaction model (“best” model; Fig. 5f; Supplementary material, Appendix S2). When the hierarchical effect was less marked (HM1, HM6) or within patch migration was high (HM2, HM3, HM6, HM7) the habitat patch model was selected (Fig. 5c, f; Supplementary material, Appendix S2). When migration probability was constant between subpatches across the entire landscape (HM4, HM8), the distance model was correctly chosen more frequently than the other two models (Supplementary material, Appendix S2).

Comparison to standard genetic statistics

Across all simulated conditions and sample sizes, we were able to detect landscape genetic structure even when F_{ST} values were minimal (0.00015–0.05, $r = 0.113$ –0.934, p-value < 0.001; Fig. 6). This result was most striking when we simulated internal patch structure, which produced more continuous genetic structure (Supplementary material, Appendix S2). Significant landscape genetic structure when F_{ST} was minimal was not the result of type I errors, which was < 0.05 for these cases.

Empirical application

We identified weak genetic structure in the black bear from N Idaho with two supported genetic clusters ($K = 2$) and $F_{ST} = 0.061$ (Pritchard et al. 2000). Raw ancestry values formed a continuous distribution from 0.0806 to 0.909 (mean = 0.482, SD = 0.227), analogous to a transition from one cover type to a second cover type (Fig. 2). We had the power to detect a landscape genetic structure, with significant correlations with all three models of landscape

genetic structure and the genetic surface (p-values 0.01–<0.0001). Out of the tested models, the patch-distance interaction model (autocorrelation with spatial dependency)

had the most support ($r=0.645$; $p\text{-value}<0.0001$). In contrast, if we had applied STRUCTURE for discrete classification of population membership with an assignment threshold of 0.75 (ancestry value), 63.7% (93/146) of individuals would be unclassified (“admixed”) indicating lack of genetic structure.

Discussion

Our approach demonstrates that it is possible to achieve one of the central goals of landscape genetics: detect the effects of the current landscape on genetic variation of a focal species, even with minimal genetic structure characteristic of fine-scale studies (Coulon et al. 2004). In addition, a genetic surfacing approach addresses some of the current limitations in landscape genetics by providing a statistically powerful, continuous representation of genetic variation. When applied to an empirical dataset our approach produced results consistent with published results. Finally, although we used discrete landscape patches for the sake of reproducibility and comparability among simulations, genetic surfacing provides a framework for modeling the effects of continuous landscape variables in empirical application.

Statistical power – data requirements

Genetic surfacing had the power to detect landscape genetic structure under conditions likely observed in extant landscapes: ≥ 5 generations post vicariance and zero to moderately high migration rates (Scribner et al. 2001, Riley et al. 2006). This is a substantial improvement over more traditional methods where up to 60 generations have been required to detect landscape effects (Keyghobadi et al. 2005, Holzhauer et al. 2006). Genetic metrics based on heterozygosity (such as F_{ST}) are less sensitive to changes in genetic variation than metrics based on allele frequency distributions (e.g. ancestry values) and could represent the effects of past process in the presence of more recent landscape change (Keyghobadi et al. 2005, Holzhauer et al. 2006).

The data required for high statistical power of genetic surfacing are practical for broad application to empirical systems, given the following considerations. First, our simulations likely underestimate microsatellite allelic diversity relative to empirical application. We randomly subsampled microsatellite loci for analysis, some of which were invariant. However, researchers generally exclude monomorphic loci and may select highly polymorphic loci.

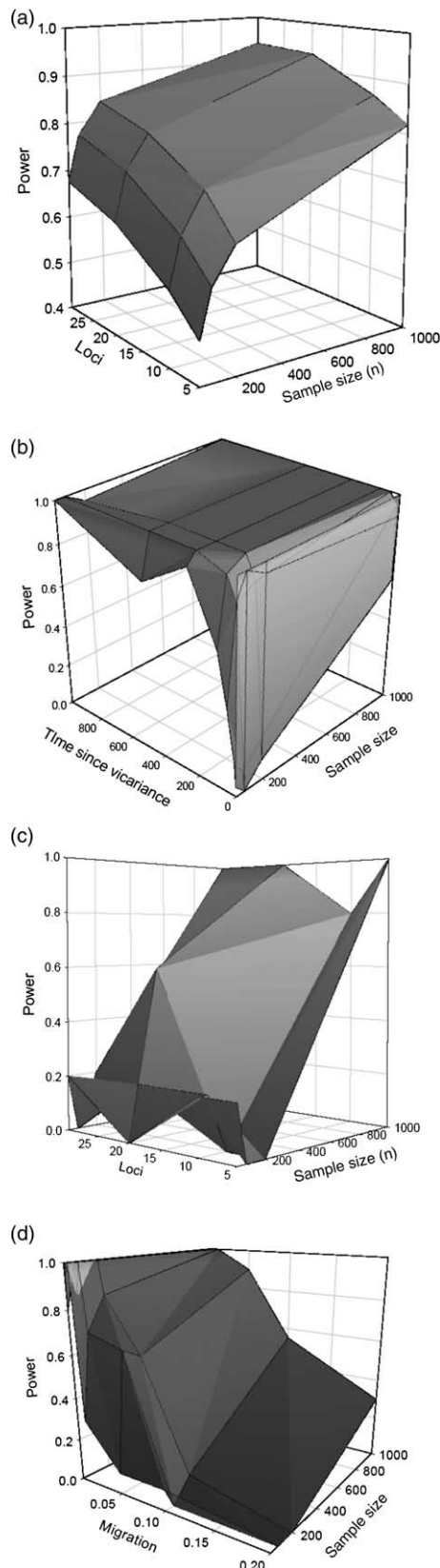


Figure 4. Power of genetic surfacing with no subsampling. (a) Statistical power of the genetic surfacing approach to detect landscape genetic structure by sample size and number of loci; (b) for all time since vicariance simulations (T), statistical power of the genetic surfacing technique to detect landscape genetic structure by time since vicariance and sample size; (c) for 10 generations post vicariance (simulation T4), statistical power of the genetic surfacing technique to detect landscape genetic structure by sample size and number of loci; (d) for all migration probability simulations (M), statistical power of the genetic surfacing technique to detect landscape genetic structure by probability of migration (migration rate) and sample size.

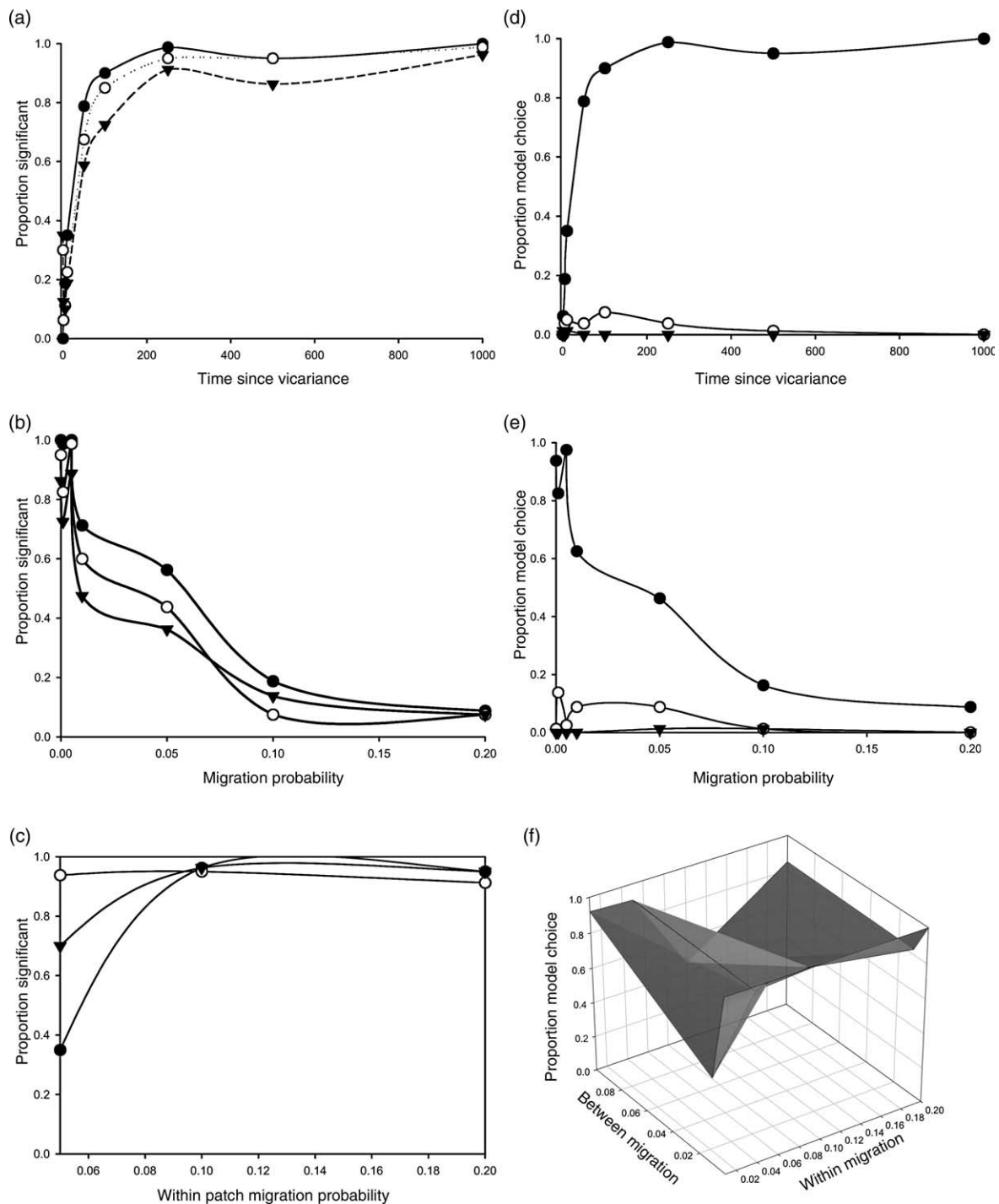


Figure 5. Accuracy of genetic surfacing. In panels (a–e) habitat patch model is shown in closed circles, distance model by open circles, and patch-distance interaction model by triangles. (a) Overall proportion of correlations between genetic surface and model of genetic structure (habitat patch, distance, patch distance interaction) significant by time since vicariance for each model. (b) Overall proportion of correlations between genetic surface and model of genetic structure significant by migration probability for each model. (c) Overall proportion of correlation between genetic surface and model of genetic structure significant by within patch migration probability. (d) Proportion of analyses for which each model of genetic structure was selected for time since vicariance. (e) Proportion of analyses for which each model of genetic structure was selected by migration probability. (f) Proportion of analyses for which the “correct” model was selected by between and within migration probabilities.

Second, adding samples to the analyses increases statistical power of genetic surfacing more quickly than adding loci (Fig. 4a). In application, it is generally easier for researchers to increase sample size than develop additional loci, which

add power in conventional population genetic application. Because we randomly sampled genotypes across the landscape, increasing sample size enhances the spatial distribution of observations improving the estimate (Legendre et al.

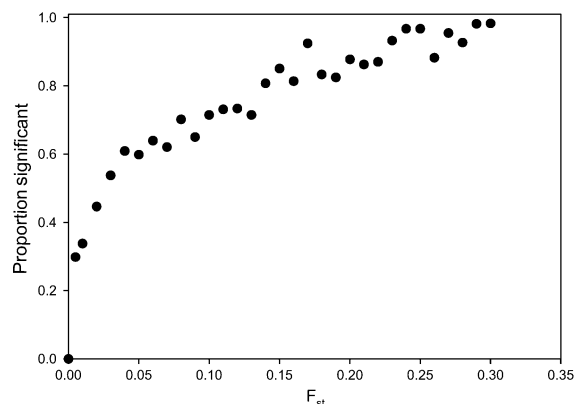


Figure 6. Proportion of Pearson's correlation coefficients between the genetic surface and simulated landscape model significant (y) by average F_{ST} value (x) across all analyses (simulation conditions, number of samples, number of loci). We can detect significant landscape process even when genetic structure is below the level generally considered biologically significant (<0.05).

2002). Third, the absolute number of samples per population does not notably increase (20–30 samples) as population size decreases or populations become unequal in size (Supplementary material, Appendix S2). Finally, more complex spatial sampling designs than the random sample employed in this study may reduce the number of samples required for equivalent power (Rempel and Kushneriuk 2003).

Accuracy and process identification

Genetic surfacing as presented in this paper is accurate and robust. The low type I error rates demonstrate that significant relationships between the genetic surface and the three landscape genetic models are not artifacts of the simulation conditions, use of ancestry values, or how landscape rasters were calculated. In addition, we show that landscape genetic structure can be distinguished from random effects and is robust to a misidentification of K . However, genetic surfacing is not dependent on STRUCTURE ancestry values used in this demonstration. Any genotypic point data calculated independently from space could be used for a surface interpolation including other clustering algorithms (Corander et al. 2003), proportion of shared alleles, or other point estimates of genetic diversity.

Although the genetic surfacing approach is accurate, misidentification of the process generating observed genetic structure is possible when multiple models are not considered. We demonstrate that distance can be significant due to “false” autocorrelation (Legendre 1993). That is, genetic surfacing detected a significant distance relationship when observations were not spatially autocorrelated but rather dependent on some other variable, which is itself autocorrelated. For example, although genetic structure was simulated solely as a product of habitat patch in landscape A (Fig. 1A), significant correlation coefficients between the genetic surface and the distance model are present due to autocorrelation of habitat patches ($I=0.98$, $p<0.001$). However, when we compared multiple models, habitat patch explained more variation than distance or patch-

distance interaction. This demonstrates that genetic surfacing can differentiate between spatial autocorrelation (i.e. isolation-by-distance) and spatial dependency (i.e. genetic structure that is dependent on an autocorrelated variable such as habitat patch) (Fortin and Dale 2005). This result emphasizes the value of incorporating landscape variables in broader population genetics research and the importance of testing multiple hypotheses of genetic structure.

Genetic surfacing did not always detect patch-distance interaction as the correct condition for data generated under this model (Supplementary material, Appendix S2). Distance was only identified if within patch migration was restricted (probability of migration = 0.05) but not extremely different from the migration probability between habitat patches (Fig. 5c). There are two principal reasons for this limitation. First, genetic divergence with high levels of migration within patches (especially with low levels of migration between patches) is typically low. Second, in the stepping-stone simulations with internal patch structure, the Pearson's correlation may have low power to distinguish between the three landscape models (distance, habitat patch, patch-distance interaction). Since genetic surfacing identified hierarchical structure when the simulated hierarchical genetic structure was most distinct (HM5), it is likely genetic signal that limits model choice more than more than spatial signal. Significance testing via a randomization method (Gardner et al. 1987) may improve model choice.

Comparison to standard genetic statistics

Recent work suggests low values of F_{ST} may be statistically, but not necessarily biologically, significant when using highly polymorphic genetic markers such as microsatellites (Hedrick 2005, Waples and Gaggiotti 2006). With high power attained with highly polymorphic loci, it is possible to detect significant subdivision (i.e. F_{ST} is significantly >0) when the actual value of the statistic is quite low (e.g. $F_{ST} < 0.05$) and the implied level of gene flow is quite high. However, our continuous genetic surfacing approach suggests ecologically relevant landscape effects may be present even when $F_{ST} < 0.05$ (Fig. 6). Because the genetic surface is a spatially continuous measure of genetic structure, it reflects population structure not captured in a global summary such as F_{ST} . Thus, analyzing the genotypic data in a spatially informed manner and using spatial analysis methods increases the power to explain how landscape features influence population genetic structure relative to standard methods.

Empirical application

In application to an empirical dataset, we demonstrate the ability to detect landscape genetic structure with a realistic amount of data (146 samples, 9 loci) and select an ecologically reasonable model using a dataset with minimal global genetic structure. In analysis of this dataset, Cushman et al. (2006) tested multiple hypotheses of genetic differentiation, including isolation-by-distance, barrier (valley), and landscape resistance. They found the landscape resistance model had the most statistical support of their tested

models. We tested three coarser-scale models of landscape genetic structure consistent with our simulations: distance (analogous to the isolation-by-distance model), habitat patch (analogous to the barrier model), and patch-distance interaction. We found a habitat patch-distance interaction to have the most support. Of our tested models, this is the only model that incorporates distance and habitat influences, both of which Cushman et al. (2006) incorporate in the landscape resistance model. Low elevation, water, and non-forest are present in the “unsuitable habitat” of our patch model and have high associated resistance in the landscape resistance model. In addition, the landscape resistance model incorporates a measure of autocorrelation by calculating cumulative costs between pairs of observations.

Our analysis of the black bear dataset was a substantial improvement over a standard population classification approach. If we used a conventional application of STRUCTURE and classified individuals into discrete “populations”, the number of unclassified individuals (63.7%) would lead to the conclusion of little structure in these data. By treating genetic structure as a continuous surface, and thus the ability to represent gradients of genetic structure, we found an increased sensitivity to spatially dependent relationships compared to a discrete classification of black bears. Finally, although we used a continuous representation of genetic variation, we classified landscape variation into discrete patches to be consistent with the simulated conditions. In future genetic surfacing applications, continuous representation of both landscape and genetic variation would be more powerful allowing for both genetic and environmental gradients (Shimatani and Kubota 2004, McGarigal and Cushman 2005).

Future directions and conclusions

Genetic surfacing provides a common statistical framework for population genetics, spatial ecology, and spatial statistics for data analysis. In future applications, the wide range of available multivariate spatial statistical techniques can be applied with genetic surfacing (Haining 2003). These multivariate models will allow researchers to quantify the effect of multiple landscape variables simultaneously for a more complete understanding of how ecological variables affect species’ distributions. In addition, the topology of the genetic surface itself may provide meaningful information. For example, genetic surfaces can be analyzed in a similar manner to other surfaces to quantify topographic complexity (Mucina et al. 1991), spatial trend (Lichstein et al. 2002), or identify areas of discontinuity and their associated landscape variables.

Application of more sophisticated surfacing methods than IDW, implemented for reproducibility and comparability among simulations, may increase the power of genetic surfacing. This is especially true in the presence of complex genetic structure resulting from multiple interacting landscape processes. Alternative methods that can produce a surface incorporate localized variation in the data such as empirical semivariogram/kriging models (Cressie 1986, Yfantis et al. 1987) or can apply more complex tension and smoothing parameters in spline models (Mitasova and Jaroslav 1993). In addition, alternative methods

allow for the inclusion of an estimate of uncertainty in the original point locations. Given multiple locations or error in geographic positioning system (GPS) locations, kernel density estimates can give a probability density function of location over a range (Wand and Jones 1995). However, it is important to understand that surfacing methods estimate a response variable across the extent and assume a well-distributed spatial sample (Tobler 1979, Lam 1983) and point data could be used directly without surface interpolation (Boots and Getis 1988, Diggle 2003). In addition, due to the complexity of data collected in natural systems, low to moderate correlations between genetic structure and spatial variables are possible by chance alone. Simulation or neutral landscape tests may be necessary to establish the significance of results (Isaaks and Srivastava 1989, Lancaster 2006).

In all of our analyses, we were able to use $K=2$ and therefore construct a single genetic surface representing membership in both genetic clusters. Although STRUCTURE is not the only potential implementation of genetic surfacing, in future application based on STRUCTURE there will be cases where more than two genetic clusters are supported. Several approaches may be employed in cases where $K > 2$. A genetic surface, and model of landscape genetic structure, could be constructed for each of $K-1$ clusters. This may be extremely useful in cases where researchers suspect different processes explain genetic variation for different genetic clusters across the landscape. Alternatively, genetic surfacing could be applied with methods that allow for multiple response variables. Fuzzy set theory allows for the response variable to have proportional membership in multiple classes, which could be modeled as a single response variable (Bosserman and Ragade 1982, Metternicht 2003).

Genetic surfacing will help fill a gap between population genetics and spatial ecology. The approach integrates landscape and genetic analyses to allow testing of multiple ecologically relevant hypotheses (autocorrelation, spatial dependency, or an interaction) without the necessity of analyzing pairwise distance data. In addition, this study demonstrates that significant distance relationships may be the result of spatial dependency or “false” autocorrelation (Legendre et al. 2002), emphasizing the importance of considering landscape measures in gene flow studies. Analysis of genetic surfaces could be readily applied to quantify the impact of ecological (e.g. temperature, moisture, natural barriers, elevation) and/ or anthropogenic (e.g. roads, development, farming, forest management) variables on genetic structure. Understanding of the ecological and spatial behavior of a population is necessary for designing recovery plans, conserving corridors and understanding potential new population threats.

Acknowledgements – We would like to thank the following people for helpful comments and discussions: J. Baumsteiger, R. Dezzani, C. Goldberg, J. Kerby, M. Morgan, A. Robinson, S. Spear, C. Steele, T. Rice, and three anonymous reviewers. In addition, we thank D. Borcard and N. Crookston for assistance in modifying the MODTTEST code, D. Hall for Perl code to extract output values for analysis, and M. Schwartz and C. Pillgrim for work on the empirical black bear data set. Funding for this research was provided by: support under National Science Foundation award number DEB-0606458 to AS and MAM, King Fellowship (MAM), Washington State Univ. Zoology Alumni Scholarship

(MAM), EPA-STAR Fellowship (#FP-453 916695; MAM), and National Science Foundation award number DEB-0548415 to AS.

References

- Arnaud, J.-F. 2003. Metapopulation genetic structure and migration pathways in the land snail *Helix aspersa*: influence of landscape heterogeneity. – *Landscape Ecol.* 18: 333–346.
- Balloux, F. 2001. A computer program for the simulation of population genetics. – *J. Hered.* 92: 301–302.
- Bockelmann, A. C. et al. 2003. Habitat differentiation vs. isolation-by-distance: the genetic population structure of *Elymus athericus* in European salt marshes. – *Mol. Ecol.* 12: 505–515.
- Boots, B. and Getis, A. 1988. Point pattern analysis. – Scientific Geography Series, Vol. 8, Sage Publ.
- Bosserman, R. W. and Ragade, R. K. 1982. Ecosystem analysis using fuzzy set theory. – *Ecol. Model.* 16: 191–208.
- Castellano, S. and Balletto, E. 2002. Is the partial mantel test inadequate? – *Evolution* 56: 1871–1873.
- Corander, J. et al. 2003. Bayesian analysis of genetic differentiation between populations. – *Genetics* 163: 367–374.
- Coulon, A. et al. 2004. Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual-based approach. – *Mol. Ecol.* 13: 2841–2850.
- Cressie, N. 1986. Kriging nonstationary data. – *J. Am. Stat. Assoc.* 81: 625–634.
- Cressie, N. A. C. 1993. Statistics for spatial data. – Wiley.
- Cushman, S. A. et al. 2006. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. – *Am. Nat.* 168: 486–499.
- Diggle, P. J. 2003. Statistical analysis of spatial point patterns. – Academic Press.
- Dutilleul, P. 1993. Modifying the t-test for assessing the correlation between two spatial processes. – *Biometrics* 49: 305–314.
- Epperson, B. K. 2003. Geographical genetics. – Princeton Univ. Press.
- Evanno, G. et al. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. – *Mol. Ecol.* 14: 2611–2620.
- Fortin, M.-J. and Payette, S. 2002. How to test the significance of the relation between spatially autocorrelated data at the landscape scale: a case study using fire and forest maps. – *Ecoscience* 9: 213–218.
- Fortin, M.-J. and Dale, M. 2005. Spatial analysis: a guide for ecologists. – Cambridge Univ. Press.
- Fotheringham, A. S. et al. 2002. Geographically weighted regression: the analysis of spatially varying relationships. – Wiley.
- Gardner, R. H. et al. 1987. Neutral models for the analysis of broad-scale landscape pattern. – *Landscape Ecol.* 1: 19–28.
- Haining, R. 2003. Spatial data analysis: theory and practice. – Cambridge Univ. Press.
- Hedrick, P. 2005. “Genetic restoration”: a more comprehensive perspective than “genetic rescue”. – *Trends. Ecol. Evol.* 20: 109.
- Holderegger, R. and Wagner, H. 2006. A brief guide to landscape genetics. – *Landscape Ecol.* 21: 793–796.
- Holzhauser, S. et al. 2006. Effect of historic landscape change on the genetic structure of the bush-cricket *Metrioptera roeseli*. – *Landscape Ecol.* 21: 891–899.
- Isaaks, E. H. and Srivastava, R. M. 1989. An introduction to applied geostatistics. – Oxford Univ. Press.
- Keyghobadi, N. et al. 2005. Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassium smintheus*, vary with landscape connectivity. – *Mol. Ecol.* 14: 1897–1909.
- Lam, N. 1983. Spatial interpolation methods: a review. – *Am. Cartogr.* 10: 129–149.
- Lancaster, J. 2006. Using neutral landscapes to identify patterns of aggregation across resource points. – *Ecography* 29: 385–395.
- Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – *Ecology* 74: 1659–1673.
- Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. – *J. Stat. Comput. Sim.* 67: 37–73.
- Legendre, P. et al. 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. – *Ecography* 25: 601–615.
- Lichstein, J. W. et al. 2002. Spatial autocorrelation and autoregressive models in ecology. – *Ecol. Monogr.* 72: 445–463.
- Manel, S. et al. 2003. Landscape genetics: combining landscape ecology and population genetics. – *Trends Ecol. Evol.* 18: 189–197.
- McGarigal, K. and Cushman, S. 2005. The gradient concept of landscape structure. – In: Weins, J. and Moss, M. (eds), Issues and perspectives in landscape ecology. Cambridge Univ. Press, pp. 112–119.
- McRae, B. H. 2006. Isolation by resistance. – *Evolution* 60: 1551–1561.
- McRae, B. H. and Beier, P. 2007. Circuit theory predicts gene flow in plant and animal populations. – *Proc. Nat. Acad. Sci. USA* 104: 19885–19890.
- Metternicht, G. I. 2003. Categorical fuzziness: a comparison between crisp and fuzzy class boundary modeling for mapping salt-affected soils using Landsat TM data and a classification based on anion ratios. – *Ecol. Model.* 168: 371–389.
- Mitasova, H. and Jaroslav, H. 1993. Interpolation by regularized spline with tension: II. Application to terrain modeling and surface geometry analysis. – *Math. Geol.* 25: 657–669.
- Mucina, L. et al. 1991. Trend surface analysis and splines for pattern determination in plant communities. – In: Fedi, E. and Orloci, L. (eds), Computer assisted vegetation analysis. Kluwer, pp. 355–371.
- Pritchard, J. K. et al. 2000. Inference of population structure using multilocus genotype data. – *Genetics* 155: 945–959.
- Raufaste, N. and Rousset, F. 2001. Are partial mantel tests adequate? – *Evolution* 55: 1703–1705.
- Rempel, R. S. and Kushneriuk, R. S. 2003. The influence of sampling scheme and interpolation method on the power to detect spatial effects of forest birds in Ontario (Canada). – *Landscape Ecol.* 18: 741–757.
- Riley, S. P. D. et al. 2006. A southern California freeway is a physical and social barrier to gene flow in carnivores. – *Mol. Ecol.* 15: 1733–1741.
- Rousset, F. 2002. Partial mantel tests: reply to Castellano and Balletto. – *Evolution* 56: 1874–1875.
- Scribner, K. et al. 2001. Environmental correlates of toad abundance and population genetic diversity. – *Biol. Conserv.* 98: 201–210.
- Shimatani, K. and Takahashi, M. 2003. On methods of spatial analysis for genotyped individuals. – *Heredity* 91: 173–180.
- Shimatani, K. and Kubota, Y. 2004. Spatial analysis for continuously changing point patterns along a gradient and its application to an *Abies sachalinensis* population. – *Ecol. Model.* 180: 359–1269.
- Slatkin, M. and Arter, H. E. 1991. Spatial autocorrelation methods in population-genetics. – *Am. Nat.* 138: 499–517.
- Sokal, R. R. et al. 1997. A simulation study of microevolutionary inferences by spatial autocorrelation analysis. – *Biol. J. Linn. Soc.* 60: 73–93.
- Sokal, R. R. et al. 1998. Local spatial autocorrelation in biological variables. – *Biol. J. Linn. Soc.* 65: 41–62.

- Spear, S. F. et al. 2005. Landscape genetics of the blotched tiger salamander (*Ambystoma tigrinum melanostictum*). – Mol. Ecol. 14: 2553–2564.
- Storfer, A. et al. 2007. Putting the “landscape” in landscape genetics. – Heredity 98: 128–142.
- Tobler, W. R. 1979. Smooth pycnophalactic interpolation for geographical regions. – J. Am. Stat. Assoc. 74: 519–530.
- Wagner, H. H. and Fortin, M.-J. 2005. Spatial analysis of landscapes: concepts and statistics. – Ecology 86: 1975–1987.
- Wand, M. P. and Jones, M. C. 1995. Kernel smoothing. – Chapman and Hall.
- Waples, R. S. and Gaggiotti, O. E. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. – Mol. Ecol. 15: 1419–1439.
- Yfantis, E. A. et al. 1987. Efficiency of kriging estimation for square, triangular, and hexagonal grids. – Math. Geol. 19: 183–205.
- Zar, J. H. 1999. Biostatistical analysis. – Simon and Schuster.

Download the Supplementary material as file E5428 from
<www.oikos.ekol.lu.se/appendix>.